建模流程、记录及输出物（用于项目进度跟踪以及建模方向调整）：

1. 模型定义（most important）

* 目标变量是什么？(y)
* 可获取原始数据有哪些？(X)
* 模型逻辑的合理性在哪儿？(X和y之间的关系)

1. 原始数据获取

* 获取时间
* 存储形式
* 原始字段中英文对照表（包括解释、备注等）
* 码表（针对原始离散型字段的字符编码与真实含义之间的映射关系）
* 数据获取初步方案：比如纯文本格式如何转换、数据库存储如何获取、数据量太大如何分群分组等
* 原始数据字段分析：
  + 是否有唯一主键，每个字段是否与提取格式相符
  + 样本量、字段数量
  + 字段类型（离散型／连续型）
  + 字段取值范围（离散型：取值种类及比例，与码表对比／连续型：取值范围及分布）
  + 字段缺失值数量及比例、异常值数量及比例
* 原始数据字段初步清洗方案（为后续的模型字段计算做准备），比如：连续型是否应做离散化，离散型如何编码（不在码表中的如何处理？），缺失如何处理（当作新类别？简单均值中位数填补？模型填补？）以及其他特殊处理问题
* 原始数据获取及清洗方案实现（程序代码编写）

1. 数据预处理（包括生成指标变量）

* 模型字段中英文对照表（包括解释、备注等）
* 模型字段计算方案（包括计算过程中遇到的缺失值、异常值等如何处理）
* 模型字段计算过程实现（程序代码编写）
* 模型字段分析：
  + 样本量、字段数
  + 字段类型（离散型／连续型）
  + 字段取值范围（离散型：取值种类及比例／连续型：取值范围及分布）
  + 字段缺失值数量及比例、异常值数量及比例
  + 字段与目标变量之间的分析图（单变量分析，不同分布的变量用多个模型去做）
* 模型字段初步入模方案，比如：连续型是否应做离散化、是否准备取log、是否需要标准化，离散型如何编码，缺失如何处理（当作新类别？简单均值中位数填补？模型填补？）以及其他特殊处理问题

1. 模型构建

* 模型字段、模型目标变量、建模样本描述（抽样？全量？）
* 算法选择、参数调整方案，如何进行效果评估
* 所有模型尝试结果日志记录（时间、操作人、入模字段处理方式、算法、参数、评估指标、备注（比如产出物较多、以文件存储时，该条操作记录对应的输出结果对应目录））

子模型较多时，可以尝试日志记录，方便汇报、自己理解、效果改进

1. 模型输出及分析

* 模型结果输出（比如概率值、拟合值、聚类中心点、异常排序、图等）
* 模型结果评估（比如概率分布、准确率、精准度、召回率、AUC、KS、ROC曲线、图等）
* 模型结构输出（比如回归系数、二叉树结构、变量重要度、图等）
* 模型解释与合理性（需要研究每个变量与模型输出结果之间的关系，比如概率得分高低的原因）
* 模型应用与落地方案（如何进行业务提升？如何与IT部门对接？如何推广？如何产品化？）